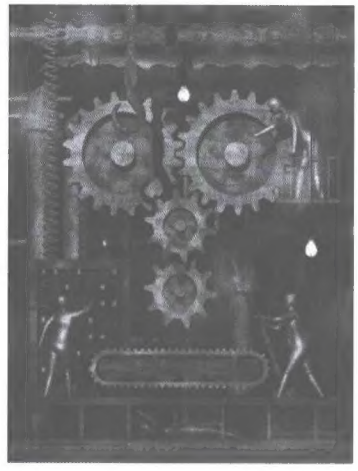
■กฤษณะ ไวยมัย ชิดชนก ส่งศิริ และ ธนาวินท์ รักธรรมานนท์วรสุชีพ

มารู้จักกับ Data Mining



การสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่มาก (Knowledge Discovery from very large Databases : KDD) หรือที่เรียกกัน Data Mining เป็นสาขาหนึ่งในวิทยาศาสตร์คอมพิวเตอร์ที่กำลังได้รับความสนใจอย่างสูงในปัจจุบัน ด้วยเทคนิคของ KDD ข้อมูลขนาดใหญ่จะถูกวิเคราะห์และสืบค้นความรู้สิ่งที่สำคัญออกมารวบรวมและจัดเก็บให้อยู่ในรูปฐานความรู้ (Knowledge Base) เพื่อใช้สำหรับการสืบค้นสิ่งที่ต้องการซึ่งไม่สามารถสืบค้นได้จากวิธีการของระบบจัดการฐานข้อมูล (DBMS) โดยทั่วไป เช่น การวิเคราะห์หาความสัมพันธ์ของข้อมูล หรือการทำนายปรากฏการณ์ต่าง ๆ ของข้อมูลที่กำลังจะเกิดขึ้น เทคนิคต่าง ๆ เหล่านี้สามารถนำไปใช้ให้เกิดประโยชน์ได้ในหลาย ๆ สาขา

วิวัฒนาการของเทคโนโลยีฐานข้อมูล

ในปี 1960 เทคโนโลยีฐานข้อมูลได้เริ่มพัฒนามาจาก file processing พื้นฐาน การค้นคว้าและพัฒนาระบบฐานข้อมูลมีมาเรื่อย ๆ จนกระทั่งปี 1970 ได้นำไปสู่การพัฒนาระบบการเก็บข้อมูลในรูปแบบตาราง (Relational Database System) มีเครื่องมือจัดการโมเดลข้อมูล และมีเทคนิคการใช้อินเด็กซ์และการบริหารข้อมูล นอกจากนี้ผู้ใช้ยังได้รับความสะดวกในการเข้าถึงข้อมูลโดยการใช้ภาษาในการเรียกข้อมูล (Query Language) เทคโนโลยีฐานข้อมูลตั้งแต่ปี 1980 ได้เริ่มมีการปรับปรุงและพัฒนาในการหาระบบการจัดการที่มีศักยภาพมากขึ้น ความก้าวหน้าในเทคโนโลยีฮาร์ดแวร์ใน 30 ปีที่ผ่านมา ได้นำไปสู่การจัดเก็บข้อมูลจำนวนมากที่มีความซับซ้อนได้อย่างมีประสิทธิภาพเพิ่มขึ้น

รู้จักกับดาต้าไมน์นิง และการประยุกต์ใช้ ดาต้าไมน์นิง กับงานด้านต่าง ๆ

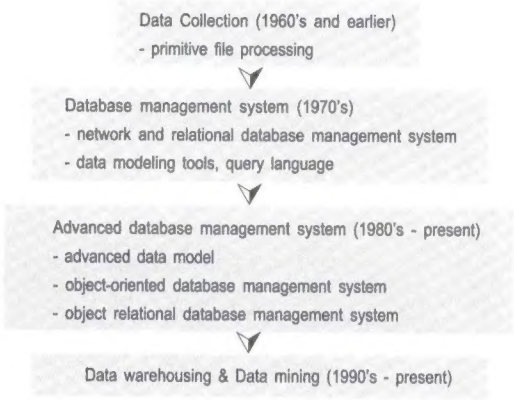
ข้อมูลในปัจจุบันสามารถจัดเก็บได้ในหลายรูปแบบ แตกต่างกันทั้งระบบปฏิบัติการ หรือการจัดเก็บฐานข้อมูล ซึ่งการนำข้อมูลทั้งหมดมารวมและจัดเก็บไว้ในรูปแบบเดียวกันเรียกว่า data warehouse เพื่อความสะดวกในการจัดการต่อไป ซึ่งเทคโนโลยี Data Warehouse รวมไปถึง Data Cleansing, Data Integration และ On-Line Analytical Processing (OLAP) ที่เป็นเทคนิคการวิเคราะห์ข้อมูลในหลาย ๆ มิตินั้นได้เกิดขึ้นมาตามลำดับ

การละเลยข้อมูล ควบคู่ไปกับการขาดเครื่องมือวิเคราะห์ข้อมูลที่มีศักยภาพ นำไปสู่คำสถานการณ์ที่ว่า "ข้อมูลมาก แต่ความรู้น้อย" (data rich but information poor) การเติบโตขึ้นอย่างรวดเร็วของข้อมูลจำนวนมากที่สะสมไว้ในฐานข้อมูลขนาดใหญ่มากซึ่งเกินกว่าที่กำลังคนจะสามารถจัดการได้ เป็นผลทำให้มีความจำเป็นที่ต้องมีเครื่องมือที่ช่วยในการวิเคราะห์ข้อมูลและหาความเป็นไปได้ของข้อมูลทั้งหมดที่เป็นประโยชน์ออกมา ก็คือ Data Mining ซึ่งได้แพร่หลายในด้านธุรกิจในยุคปัจจุบัน

ทำไมจึงต้องมี Data Mining

ปัจจัยที่ทำให้ดาตาไมน์นิงเป็นที่ได้รับความสนใจอย่างสูงมีดังต่อไปนี้

- จำนวนและขนาดข้อมูลขนาดใหญ่ถูกผลิตและขยายตัวอย่างรวดเร็ว การสืบค้นความรู้จะมีความหมายก็ต่อเมื่อฐานข้อมูลที่ใช้มีขนาดใหญ่มาก ปัจจุบันมีจำนวนและขนาดข้อมูลขนาดใหญ่ที่ขยายตัวอย่างรวดเร็ว โดยผ่านทางอินเทอร์เน็ต ดาวเทียม และแหล่งผลิตข้อมูลอื่น ๆ เช่น เครื่องอ่านบาร์โค้ด เครดิตการ์ด อีคอมเมิร์ซ เป็นต้น



รูปที่ 1 วิวัฒนาการเทคโนโลยีฐานข้อมูล

• ข้อมูลถูกจัดเก็บเพื่อนำไปสร้างระบบการสนับสนุนการตัดสินใจ เพื่อเป็นการง่ายต่อการนำข้อมูลมาใช้ในการวิเคราะห์เพื่อการตัดสินใจ ส่วนมากข้อมูลจะถูกจัดเก็บแยกมาจากระบบปฏิบัติงาน (Operational Systems) โดยจัดอยู่ในรูปของคลังหรือเหมืองข้อมูล (Data Warehouse) ซึ่งเป็นการง่ายต่อการนำเอาไปใช้ในการสืบค้นความรู้

• ระบบคอมพิวเตอร์สมรรถนะสูงมีราคาต่ำลง เทคนิคดาตาไมน์นิงประกอบไปด้วยอัลกอริธึมที่มีความซับซ้อนและความต้องการการคำนวณสูง จึงจำเป็นต้องใช้งานกับระบบคอมพิวเตอร์สมรรถนะสูง ปัจจุบันระบบคอมพิวเตอร์สมรรถนะสูงมีราคาที่ต่ำลง พร้อมด้วยเริ่มมีเทคโนโลยีที่นำเครื่องไมโครคอมพิวเตอร์จำนวนมากมาเชื่อมต่อกันโดยเครือข่ายความเร็วสูง (PC Cluster) ทำให้ได้ระบบคอมพิวเตอร์สมรรถนะสูงในราคาถูก

• การแข่งขันอย่างสูงในด้านอุตสาหกรรมและการค้า เนื่องจากปัจจุบันมีการแข่งขันอย่างสูงในด้านอุตสาหกรรมและการค้า มีการผลิตข้อมูลไว้อย่างมากมายแต่ไม่ได้นำมาใช้ให้เกิดประโยชน์ จึงเป็นการจำเป็นอย่างยิ่งที่ต้องควบคุมและสืบค้นความรู้ที่ถูกซ่อนอยู่ในฐานข้อมูล ความรู้ที่ได้รับสามารถนำไปวิเคราะห์เพื่อการตัดสินใจในการจัดการในระบบต่าง ๆ ซึ่งจะเห็นได้ว่าความรู้เหล่านี้ถือว่าเป็นผลิตผลอีกชิ้นหนึ่งเลยทีเดียว

Data Mining ทำอะไรได้บ้าง

การแบ่งประเภทหรือระบุถึงดาตาไมน์นิงนั้นค่อนข้างลำบาก เพราะไม่มีการแบ่งหรือระบุอย่างชัดเจน ซึ่งถ้าแบ่งตามหน้าที่หลัก จะสามารถแบ่งได้ 2 ประเภท คือ

1. Descriptive เป็นการบอกถึงลักษณะ ข้อสรุป อธิบายคุณสมบัติโดยทั่วไปในฐานข้อมูล แบ่งได้เป็น 2 แบบ คือ

• **Characterization** การหาข้อสรุปที่ถูกต้องจากข้อมูล เช่น หาลักษณะของลูกค้าที่ใช้เงินมากกว่า $1000 ต่อปี

• **Discrimination** การเปรียบเทียบข้อมูล เช่น ลักษณะลูกค้าที่ซื้อสินค้าเป็นประจำและลูกค้าที่แทบจะไม่ซื้อสินค้าเลย

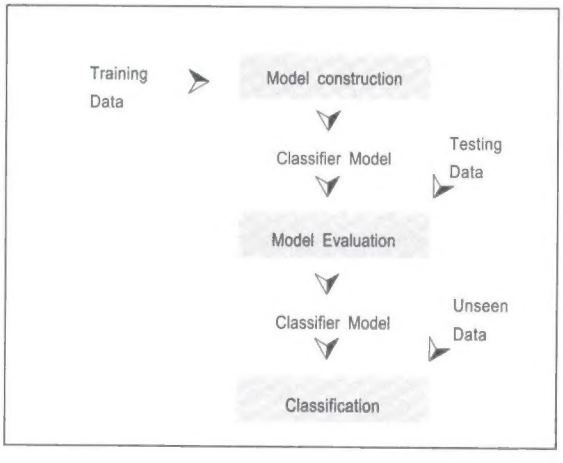
2. Predictive เป็นการวิเคราะห์หาโมเดลเพื่อทำนายแนวโน้มหรือคุณสมบัติของข้อมูลที่ยังไม่ทราบ เช่น การหายอดขายสินค้าของเดือนถัดไป

เทคนิคต่าง ๆ ของ Data Mining

1. Association rule Discovery เป็นเทคนิคหนึ่งของ Data mining ที่สำคัญ และสามารถนำไปประยุกต์ใช้ได้จริงกับงานต่าง ๆ หลักการทำงานของวิธีนี้ คือการค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่ เพื่อนำไปใช้ในการวิเคราะห์ หรือทำนายปรากฏการณ์ต่าง ๆ ตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้กับงานจริง ได้แก่ ระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติของ Amazon ข้อมูลการสั่งซื้อทั้งหมดของ Amazon ซึ่งมีขนาดใหญ่มาก จะถูกนำมาประมวลผลด้วยเทคนิค Association rule discovery เพื่อหาความสัมพันธ์ของข้อมูล คือลูกค้าที่ซื้อหนังสือเล่มหนึ่ง ๆ มักจะซื้อหนังสือเล่มใดพร้อมกันด้วยเสมอ ความสัมพันธ์ที่ได้จากกระบวนการนี้จะสามารถนำไปใช้คาดเดาได้ว่าควรแนะนำหนังสือเล่มใดเพิ่มเติมให้กับลูกค้าที่เพิ่งซื้อหนังสือจากทางร้าน ตัวอย่างเช่น buys(x, database) -> buys(x, datamining) [80%,60%] หมายความว่า เมื่อซื้อหนังสือดาต้าเบสแล้ว มีโอกาสที่จะซื้อหนังสือดาต้าไมน์นิงด้วย 60 เปอร์เซ็นต์ และมีการซื้อทั้งหนังสือดาต้าเบสและหนังสือดาต้าไมน์นิงพร้อม ๆ กัน 80 เปอร์เซ็นต์

2. Classification & Prediction

• **Classification** เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ ตัวอย่างเช่น จัดกลุ่มนักเรียนว่าดีมาก ดี ปานกลาง ไม่ดี โดยพิจารณาจากประวัติและผลการเรียน หรือแบ่งประเภทของลูกค้าว่าเชื่อถือได้หรือไม่โดยพิจารณาจากข้อมูลที่มีอยู่ กระบวนการ classification นี้แบ่งออกเป็น 3 ขั้นตอน ดังรูปที่ 2



รูปที่ 2 กระบวนการ classification

- **Model construction (Learning)** เป็นขั้นการสร้างโมเดลโดยอาศัยการเรียนรู้จากข้อมูลที่ได้กำหนดคลาสไว้เรียบร้อยแล้ว (training data) ซึ่งโมเดลที่ได้อาจแสดงในรูปของ decision tree คล้ายโครงสร้างต้นไม้ ที่แต่ละโหนดแสดง attribute แต่ละกิ่งแสดงผลในการทดสอบ และลีฟโหนดแสดงคลาสที่กำหนดไว้ ซึ่ง decision tree นี้ง่ายต่อการปรับเปลี่ยนเป็น classification rule

- **Model evaluation (Accuracy)** เป็นขั้นการประมาณความถูกต้องโดยอาศัยข้อมูลที่ใช้ทดสอบ (testing data) ซึ่งคลาสที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับคลาสที่หามาได้จากโมเดลเพื่อทดสอบความถูกต้อง

- **Model usage (Classification)** เป็นโมเดลสำหรับใช้กับข้อมูลที่ไม่เคยเห็นมาก่อน (unseen data) โดยจะทำการกำหนดคลาสให้กับออบเจ็กต์ใหม่ที่ได้มา หรือทำนายค่าออกมาตามที่ต้องการ

• **Prediction** เป็นการทำนายหาค่าที่ต้องการจากข้อมูลที่มีอยู่ ตัวอย่างเช่น หายอดขายของเดือนถัดไปจากข้อมูล หรือทำนายโรคจากอาการของคนไข้ในอดีต เป็นต้น

3. Clustering เป็นการวิเคราะห์หาคลาสให้กับข้อมูลโดยไม่รู้ชื่อคลาสมาก่อน ซึ่งต่างกับ classification และ prediction ที่วิเคราะห์หาคลาสให้กับข้อมูลโดยมีการกำหนดคลาสไว้ให้ก่อนแล้ว ตัวอย่างของ clustering เช่น ช่วยฝ่ายการตลาดในการแบ่งกลุ่มลูกค้าเพื่อที่จะพัฒนาการขายให้ตรงเป้าหมายหรือตรงตามความต้องการในแต่ละกลุ่ม

ทุกรูปแบบที่หามาได้จากการไมน์น่าสนใจทั้งหมดหรือไม่ ?

การใช้ดาต้าไมน์นิงนั้นสามารถผลิตรูปแบบออกมาได้เป็นพันหรือล้านรูปแบบ แต่มีรูปแบบเพียงส่วนน้อยเท่านั้นที่จัดว่าน่าสนใจ โดยพิจารณาจาก

1. ง่ายต่อการเข้าใจของมนุษย์
2. มีเหตุผลและถูกต้อง
3. มีประโยชน์
4. แปลกใหม่ ซึ่งรูปแบบที่น่าสนใจนี้ก็คือความรู้ (knowledge) นั่นเอง

การประยุกต์ของ Data Mining

การวิเคราะห์ผลิตภัณฑ์

- เก็บรวบรวมลักษณะและราคาของผลิตภัณฑ์ทั้งหมด สร้างโมเดลด้วยเทคนิคดาต้าไมน์นิง และใช้โมเดลในการทำนายราคาผลิตภัณฑ์ตัวอื่น ๆ

การวิเคราะห์บัตรเครดิต

- ช่วยบริษัทเครดิตการ์ดตัดสินใจในการที่จะให้เครดิตการ์ดกับลูกค้าหรือไม่
- แบ่งประเภทของลูกค้าว่ามีความเสี่ยงในเรื่องเครดิตต่ำ ปานกลาง หรือสูง
- ป้องกันปัญหาเรื่องการทุจริตบัตรเครดิต

การวิเคราะห์ลูกค้า

- ช่วยแบ่งกลุ่มและวิเคราะห์ลูกค้าเพื่อที่จะผลิตและเสนอสินค้าได้ตรงตามกลุ่มเป้าหมายแต่ละกลุ่ม
- ทำนายว่าลูกค้าคนใดจะเลิกใช้บริการจากทางบริษัทภายใน 6 เดือนหน้า

การวิเคราะห์การขาย

- พบว่า 70 เปอร์เซ็นต์ของลูกค้าที่ซื้อโทรทัศน์แล้วจะซื้อวิดีโอตามมา ดังนั้นผู้จัดการจึงควรมุ่งไปลูกค้าที่ซื้อโทรทัศน์ แล้วจึงส่งเมล์ไปยังลูกค้าเหล่านั้นเพื่อที่จะเชิญชวน หรือให้ข้อเสนอที่ดี เพื่อให้ลูกค้ามาซื้อวิดีโอในครั้งต่อไป
- ช่วยในการโฆษณาสินค้าได้อย่างเหมาะสมและตรงตามเป้าหมาย
- ช่วยในการจัดวางสินค้าได้อย่างเหมาะสม

อีคอมเมิร์ซ

- ช่วยให้เข้าใจพฤติกรรมของลูกค้า เช่น ลูกค้ามักเข้าไปที่เว็บใดตามลำดับก่อนหลัง
- ช่วยในการปรับปรุงเว็บไซท์ เช่น พิจารณาว่าส่วนใดของเวปที่ควรปรับปรุงหรือควรเรียงลำดับการเชื่อมโยงในแต่ละหน้าอย่างไรเพื่อให้สะดวกกับผู้เข้าเยี่ยมชม

สิ่งที่ท้าทายใน Data Mining

• การหาทฤษฎีใหม่ ๆ ของดาต้าไมน์นิง

• การปรับปรุงประสิทธิภาพเทคนิคต่าง ๆ ของดาต้าไมน์นิงให้ดีและเร็วมากขึ้น

• หาวิธีใหม่ ๆ ของดาต้าไมน์นิง เช่น การรวมเอาเทคนิคของดาต้าไมน์นิงหลาย ๆ อย่างเข้าด้วยกัน

• การหาเครื่องมือเพื่อสนับสนุนดาต้าไมน์นิง โดยกระบวนการดาต้าไมน์นิงควรโต้ตอบกับผู้ใช้ (interactive) ได้

• การนำดาต้าไมน์นิงไปใช้กับระบบข้อมูลที่ก้าวหน้า เช่น Spatial, มัลติมีเดีย หรือ Web-mining

ฉบับนี้เราได้รู้จักกับดาต้าไมน์นิง และการประยุกต์ใช้ดาต้าไมน์นิงกับงานด้านต่าง ๆ ไปพอสมควรแล้วนะครับ สำหรับฉบับต่อไปเราจะได้รู้ถึงวิธีการเตรียมข้อมูลเพื่อที่จะนำมาทำการไมน์ต่อไป

แหล่งอ้างอิง

-หนังสือ

- Robert Groth: DATA MINING, Prentice Hall, Inc, 1997.
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
- M. Berry and G. Linoff. Data Mining Techniques: For Marketing, Sales and Customer Support, John Wiley & Sons, 1997.

-เอกสารแนะนำ

- M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.
- T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39:58-64, 1996.

Key Web site: Knowledge Discovery Resource Center http://www.kdnuggets.com/